Reprinted from THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA, Vol. 42, No. 4, 830–843, October 1967
Copyright, 1967 by the Acoustical Society of America.
Printed in U. S. A.

On the Rôle of Formant Transitions in Vowel Recognition

B. E. F. LINDBLOM*

Department of Speech Communication, Royal Institute of Technology (KTH), Stockholm 70, Sweden

M. STUDDERT-KENNEDY†

Haskins Laboratories, New York, New York 10017

An inventory of speechlike sounds was synthesized displaying systematic variations of the rate and direction of formant transitions. These sounds were specified by a set of vowel formant patterns selected along a continuum varying from [ʊ] to [ɪ]; they were assigned to isolated, steady-state vowels, and to the points of zero rate of formant frequency change in symmetrical consonant–vowel–consonant syllables. The time variations of formant frequencies were made convex and concave by the choice of two consonantal frames: [w-w] and [j-j]. The results obtained in a series of vowel identification experiments indicate that a listener's categorization of the continuum varied as a function of the environment and the duration of the vowel. These findings suggest that, in the recognition of monosyllabic nonsense speech, the identity of a vowel is determined not solely by the formant-frequency pattern at the point of closest approach to target, but also by the direction and rate of adjacent formant transitions. In general, subjects adjusted their categorizations of the continuum in the consonantal contexts in such a way that complete transitions between loci and vowel target were not necessary: the transitions were permitted to undershoot the target frequencies for the vowel. In particular, the excursions of formants in the [w-w] syllables tended to be overestimated. Thus, there was a tendency for the categorizations to be made so as to compensate for the formant-frequency undershoot associated with vowel reduction [B. Lindblom, "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. 35, 1773–1781 (1963)]. The effects observed are discussed in terms of an active model of vowel recognition, peripheral auditory analysis, distinctive features, and previously reported observations on vowel perception.

I. PROBLEM OF VOWEL RECOGNITION

THIS study deals with an aspect of human vowel recognition. Basic to the exploration of this process is an understanding of the conditions under which the mechanisms underlying recognition operate. For the vowel system of a given language, these conditions are determined by the relation between the sounds to be decoded and the categories to be recognized. This relation, in turn, is dependent on how the acoustic realizations of vowels are constrained in the generation of speech. In this article, we begin with a brief examination of the mechanism of speech production in order to point out a problem inherent in the recognition of vowel sounds generated by this mechanism. The results of an experiment on vowel identification will then be presented and discussed in the light of this problem.

The range of acoustic shapes characteristic of vowels is limited by the constraints of speech production. For a given articulatory system, constants and physical laws set certain margins of performance beyond which the system cannot go. Within these limits, the available degrees of freedom are used in two ways. On the one hand, there are articulatory control mechanisms related to the categories and processes postulated in the formal description of a language: these mechanisms exploit the available physiological apparatus for linguistic purposes. On the other hand, there are extralinguistic mechanisms, such as those producing variations in speaking rate, vocal effort, etc., or reflecting the play of psychological variables: these mechanisms tend to perturb the former basic underlying processes. It is hypothesized herein that, in terms of the substrate of linguistically determined *articulatory control*, an utterance is organized as a succession of vocal-tract states, there being a low

* During the initial stages of this work guest researcher at the Res. Lab. Electron., MIT, Cambridge, Mass., and Haskins Laboratories, New York.

† Present address: Inter-American University, San German, P. R.

number of such states associated with each phoneme. At the level of *articulatory performance*, however, the picture is different, for the gestures invoked to actualize these states are relatively slow: They merge spatially and temporally into a continuous process that often only approximates the intended states and continually exhibits instances of coarticulation. For experimental data compatible with these ideas and similar views of speech production, see Refs. 1–8. When the control is relaxed or perturbed by extralinguistic variables, these effects become even more marked so that ellipsis and reduction may result. For syllable production, a superordinate timing mechanism is required to control the relative moments at which maneuvers towards adjacent consonant and vowel goals are initiated. Since there are physiological and mechanical limitations on the rates at which such maneuvers can be carried out, the extent of the articulatory movements within a given syllable can vary as a function of the temporal pattern of syllabic gesture initiation. A consequence of this organization for the pronunciation of a vowel is that the speech organs sometimes undershoot the articulatory target of the vowel. As the temporal proximity of adjacent gesture initiations increases, the undershoot effect observed in the vowel of, for example, a consonant–vowel–consonant (CVC) syllable tends also to increase. Acoustically, the typical vowel segment will then appear as a continuously varying event. Only rarely will its formants reach a steady state. At some point during the segment, however, they will approach the pattern corresponding to the underlying target more closely than at other instants. Formant frequencies sampled at this moment, may be considerably displaced from their target values owing to the undershoot and perturbation effects. (Refs. 2, 3, 4, and 6.)

The acoustic information on a vowel phoneme generally present in the speech wave can be stylized as a sequence of three elements: transition+pattern at point of closest approach to target+transition. We have pointed out that these elements are subject to considerable contextual modification. These modifications sometimes make the middle element ambiguous. An illustration of such ambiguity is given in Fig. 1. This figure is based on spectrographic measurements of the formant frequencies of eight long and eight short Swedish vowels pronounced by one talker. The data refer to samples at points of closest approach to target. The immediate environment of these sounds was [ɛ'd—d], preceded by a carrier phrase. The talker was instructed to attempt to vary his rate of speaking in synchrony with a timing signal presented to him over an earphone. Each vowel was repeated about 18 times at different rates. In an F_1–F_2 plot, it can be seen that the sampling procedure used causes considerable overlap of the vowel areas. Taking the third formant into account reduces this overlap somewhat, especially in the front and close vowel region, but is far from removing it completely. There is every reason to believe that, for normal conversational speech, similar data would show even greater overlap. In a language with a rich vowel system such as Swedish, the risk of ambiguity would seem to be particularly great. But in other types of languages such as Japanese, ambiguity has also been shown to occur.[9] For ambiguity in American English vowels, see Figs. 5 and 6, which are discussed later.

From the point of view of the above singie-sample specifications, the problem of vowel recognition is to recover the identity of the underlying target in the face of ambiguity and large perturbation effects. Obviously, in human recognition of speech this problem is normally

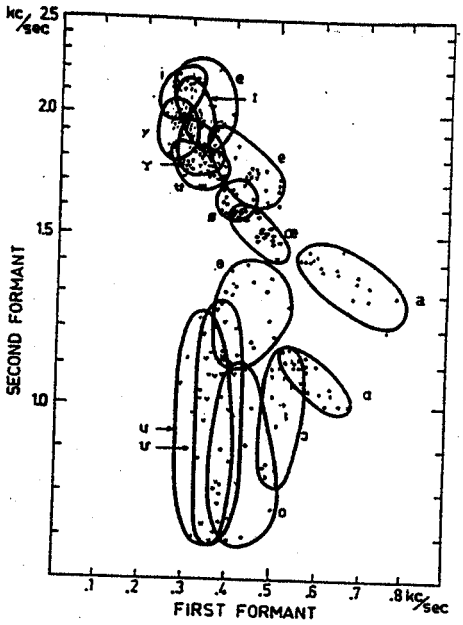

FIG. 1. Ambiguity of raw first and second formant data sampled at points of closest approach to target (Swedish).

[1] M. Halle and K. N. Stevens, "Speech Recognition: A Model and a Program for Research," IRE Trans. Inform. Theory IT-8, 155–159 (1962).

[2] K. N. Stevens and A. S. House, "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study," J. Speech & Hearing Res. 6, 111–128 (1963).

[3] B. Lindblom, "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. 35, 1773–1781 (1963).

[4] B. Lindblom, "Articulatory Activity in Vowels," Speech Transmission Lab. QPSR 2, p. 1 (1964).

[5] S. E. G. Öhman, "Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Am. 39, 151–168 (1966).

[6] K. N. Stevens, A. S. House, and A. P. Paul, "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," J. Acoust. Soc. Am. 40, 123–132 (1966).

[7] S. E. G. Öhman, "Numerical Model of Coarticulation," J. Acoust. Soc. Am. 41, 310–320 (1967).

[8] W. L. Henke, "Dynamic Articulatory Model of Speech Production Using Computer Simulation," PhD thesis, MIT, Sept. 1966.

[9] O. Fujimura and K. Ochiai, "Vowel Identification and Phonetic Contexts," J. Acoust. Soc. 35, 1889 (A) (1963).

coped with. To explain how, we might argue that a listener relies on "context." For instance, he may use the more absolutely identifiable cues in longer segments of speech to infer the identity of the reduced vowels from various phonological, syntactic, and semantic constraints. Or again, perhaps our selection of an isolated, acoustic vowel attribute gives a misleading perspective of the vowel recognition problem: Other information in the short-term acoustic context, such as the direction and rate of adjacent formant transitions, may also be important in the auditory representation of the signal and the processes of symbol assignment. It is precisely this latter possibility that is examined in the present paper.

An identification experiment was performed in order to investigate the relative rôles that formant transitions and the formant pattern at the point of closest approach to target play in vowel recognition, and to ascertain whether operations contributing to the "solution" of the vowel-recognition problem, as formulated above, are invoked at the level of monosyllabic nonsense speech. In the experiment, American English listeners were asked to identify vowel sounds presented under steady-state conditions and in symmetrical CVC syllables. The vowel-formant patterns assigned to the points of closest approach to target were selected from a continuum ranging from [ɪ] to [ʊ]. The rate and direction of the adjacent transitions were systematically varied by the choice of two consonantal frames: [w-w] and [j-j]. Two main possible outcomes of the experiment were anticipated. It was argued that a listener might categorize the vowel continuum in one of two ways.

- *Context-free case:* Categorization would be the same for the #V#, [w], and [j] contexts. As long as two stimuli had identical formant frequencies at their midpoints they would be perceptually equivalent no matter how different the adjacent formant frequency transitions. This result would indicate that sampling occurs at zero rate of change of formants, and that the measurement of the midpoint formant frequencies can be made by the auditory system irrespective of transitional context. Symbol assignment would be based on a straightforward classification of auditory patterns.

- *Context-dependent case:* Categorization would vary: The boundary between [ɪ] and [ʊ] vowels would shift as a function of context. The interpretation of this result would depend on the details of the boundary displacements.

II. STIMULUS SPECIFICATION

The stimuli were constructed, according to the above three-element model, so as to contain an initial transition, a pattern of closest approach to target, and a final transition. The basic building blocks were 20 points along a vowel-formant continuum and two sets of consonantal loci. The loci served as starting points for the initial transitional segment and the vowel patterns as the terminals of this segment. The final segment was always the mirror image of the first half of the stimulus. Thus, these syllables were of symmetrical CVC structure. At the start of the stimulus, formant frequencies remained stationary at locus frequencies for about 20 msec, then moved on to one of the vowel patterns, and from there symmetrically back to the same locus frequencies again. A 20-msec steady-state pattern at these frequencies terminated the syllable. In this way, each set of loci was combined with each vowel formant pattern.

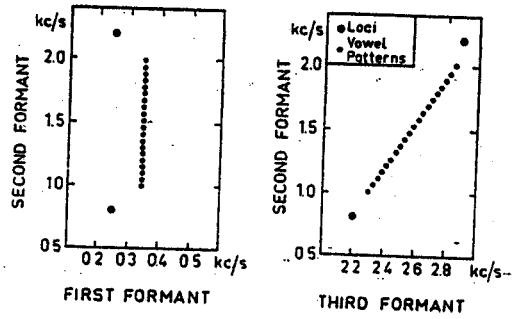

FIG. 2. Formant frequencies of vowels and consonants used to generate synthetic stimuli of #V# and CVC structure.

The vowel patterns were selected along a continuum in the F_1–F_2–F_3 space as follows:

$$F_1 = 350,$$
$$F_2 = 1000 + [(n-1)\cdot 1000]/19,$$
$$F_3 = 2300 + [(n-1)\cdot 525]/19,$$

where n represents stimulus number. Thus, F_1 is constant, while F_2 varies in linear steps of about 53 cps from 1000 to 2000 cps, and F_3 varies between 2300 and 2825 cps in steps of approximately 28 cps. Perceptually, at the extremes of this continuum, the vowel sounds approach American English [ɪ] and [ʊ]. The sets of loci chosen were a [j] set in which $F_1 = 250$, $F_2 = 2200$, $F_3 = 2900$ cps, and a [w] set where $F_1 = 250$, $F_2 = 800$, $F_3 = 2200$ cps. The information on vowels and loci is represented graphically in Fig. 2.

The relative locations of the loci and the vowel continuum were selected so as to produce initial F_2 and F_3 movements that would always exhibit either positive slopes ([w] series) or negative slopes ([j] series). The general form of the formant-frequency variation throughout the entire vowel segment was parabolic. Thus, in the [j] series, the F_2 and F_3 transitions were symmetrical parabolas, *concave* upward in a frequency–time display. The vowel continuum patterns in this set were assigned to points of zero rate of frequency change, where F_2 and F_3 reached their *minimum* values. In the [w] series the F_2 and F_3 transitions were also symmetrical parabolas; but *convex* upward. Again, the vowel patterns occurred at points

of zero rate of change, but at these points F_2 and F_3 now reached their *maximum* values. The course of F_1 was identical in the two series: It was convex upward and always reached its maximal value at the zero rate of change point. The case of zero slope of all transitions was also included to produce 20 steady-state versions of the vowel patterns. Finally, duration was introduced as a variable, the vowel segments being either 200- or 100-msec long. Twenty vowel patterns, two sets of consonant loci, one steady-state frame, and two rates give a total of $(20\cdot2\cdot2+20\cdot2)=120$ stimuli.

These syllables were synthesized on OVE II at the Department of Speech Communication (Speech Transmission Laboratory), KTH, Stockholm. The spectral shape of the source and radiation factors were simulated by a single real-axis pole at 50 cps producing a slope of -6 dB/oct. The correction for higher poles that is introduced by means of the KH and F5' circuits was at the standard settings of 4 kcps. The frequency of the fourth formant was fixed at 3500 cps and formant bandwidths as follows: $B_1=50$, $B_2=60$, $B_3=80$, and $B_4=125$ cps. A monotone fundamental frequency of 100 cps was used. The OVE II formant circuits are connected in series.[10,11]

In the stationary vowel series, formant frequencies were adjusted manually: Only the ON and OFF sets of the A_0, or voice source intensity, gate were programmed from the OVE II function generator. In the CVC syllables, two control patterns for the parameters were used to generate the stimuli, one for each consonantal environment. The appropriate frequency values for the loci and vowel patterns were obtained by calibrating the frequency range of the formant in question. This was done to improve the accuracy of frequency control. A frequency analysis of each formant circuit—repeated before each syllable was synthesized—indicated that formant frequencies were within ±5 cps of the desired frequencies. Spectrograms of each individual test stimulus were examined before the preparation of test tapes. Figure 3 presents spectrographic illustrations of some of the stimuli generated.

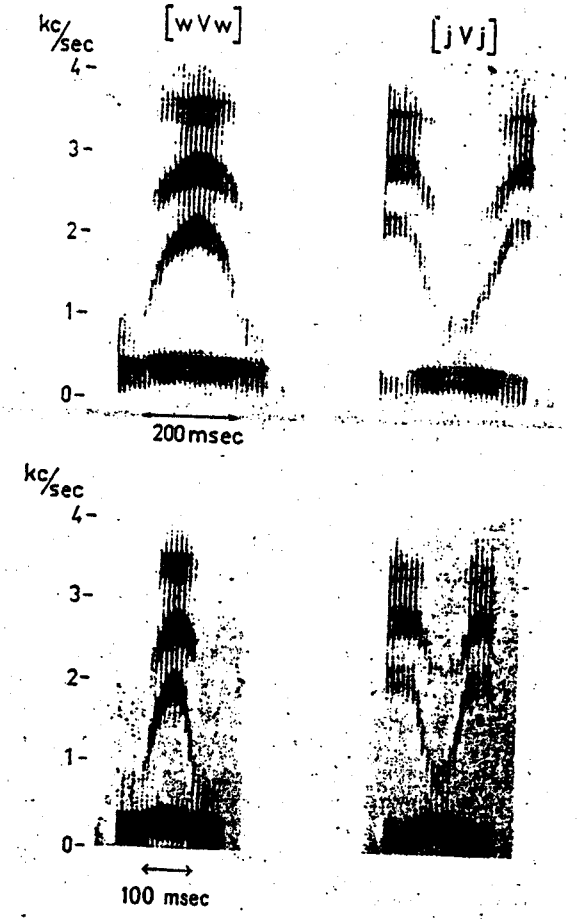

FIG. 3. Spectrographic illustrations of stimuli in the [w-w] and [j-j] series.

III. EXPERIMENTAL PROCEDURES

The method used to study the perception of the vowels was a two-alternative forced-choice identification task. Two groups of American English listeners were asked to label the vowel sounds as either [ɪ] or [ʊ]. Data are available from six subjects run at Haskins Laboratories (CA, KE, FE, KR, SO, AD), New York, and four subjects run at KTH, Stockholm (HW, DW, MS, JH). All listeners were native speakers of American English, of fairly homogeneous dialectal backgrounds and without obvious speech or hearing defects. With the exception of one subject (JH), they had had no previous experience in auditory and psychophysical tasks. No one was informed what the test was about either before or during the course of the testing. Subjects were paid for their work. The tests were distributed over a period of 10 days with one session per day, and the listening material assigned to each session lasted, at most, 50 min, exclusive of interruptions and breaks. The stimuli were arranged in four major groups: slow [w] and [j] series, fast [w] and [j] series, slow #V# series and fast #V# series. Four different randomizations of the stimuli in each series were used. Each series began with five buffer stimuli that were not used in subsequent processing of the data. On a given test tape, there were four 3–4 min series, one representative from each major group; the order of the various CVC and #V# tests was counterbalanced across the tapes. The total number of responses to each individual stimulus obtained from each listener was at least 15. On the average, each observer made

[10] G. Fant, "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Tech. No. 1, 15, 3–108 (1959).
[11] G. Fant, J. Mártony, U. Rengman, and A. Risberg, "OVE II Synthesis Strategy," in *Proceedings of the Speech Communication Seminar Stockholm 1962* (Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, 1963), Vol. 2, paper F5.

about 2000 responses in all. The New York and Stockholm subjects were run under almost identical conditions. They were seated in quiet rooms or booths and listened to the stimuli over earphones. To estimate the absolute level of the stimuli at the subjects' ears, the beginning of each tape had a 1000-cps sinusoidal signal that had been recorded at the level of the maximum reading for the vowel stimuli on the occasion of the synthesis. The gain of the tape-recorder playback for this tone was adjusted to give a voltage across the earphones equivalent to 80 dB *re* 0.0002 dyn/cm². The instructions read to the listeners were as follows:

"This is an experiment in speech perception. You are going to hear a sequence of vowel sounds. You are asked to identify each one as either [ɪ], as in 'bit,' or [ʊ], as in 'book.' In the first two tests, you will hear each vowel placed between two semiconsonantal sounds. Sometimes you will hear them between two [w]'s, as in [wɪw] or [wʊw]; other times you will hear them between two [j]'s, as in [jɪj] or [jʊj]. But whatever their context, you are asked to identify only the vowel sound in the middle. If you think the vowel is more like the [ɪ] in 'bit,' write *i* on your answer sheet opposite the appropriate stimulus number; if you think the vowel is more like the [ʊ] in 'book,' write *oo* on your answer sheet opposite the appropriate stimulus number.

"There are 45 stimuli in each of the first two tests with a 4-second interval between stimuli and a 9-second interval after the 15th, 25th, and 35th stimuli. So, each test will last for about 3 or 4 minutes. We will pause for a moment between tests and you will be warned when the second one is going to begin.

"In the next two tests you will again hear the vowels, but this time you will hear them alone without any surrounding context. Again you are asked to identify each vowel as either [ɪ], as in 'bit,' or [ʊ], as in 'book,' and to record your judgment opposite the appropriate stimulus number on your answer sheet.

"There are 25 stimuli in each of the next two tests with a 4-second interval between stimuli. So each test will last about two minutes. We will pause for a moment between tests and you will be warned when the second one is going to begin."

These instructions were read to the participants before the first four tests with appropriate pauses for writing the symbols and key words on the blackboard, and for questions and answers. Between every four tests, a longer break of about 5 min was normally taken. At subsequent sessions, the full instructions were not repeated, but a summary was given: "The tests today are similar to those of previous days. Remember, you are asked to identify each vowel as either [ɪ], as in 'bit,' or [ʊ], as in 'book.'"

IV. EXPERIMENTAL RESULTS I

A. Slow Stimuli (Vowel Duration: 200 msec)

Data were plotted for the observers individually in the form shown in Fig. 4. This figure shows the percentage of [ɪ] responses (dots) and [ʊ] responses (crosses) as a function of the stimulus continuum. Low-stimulus numbers refer to vowel patterns with low F_2 and F_3 values (see Section on stimulus specification, above). Reading from the top, the graphs show the

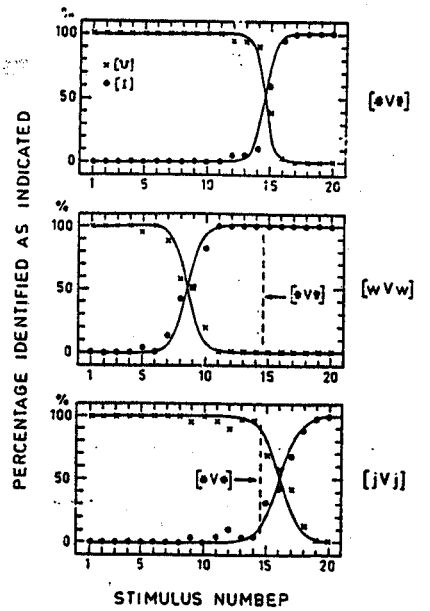

FIG. 4. The percentage of [ɪ] responses (dots) and [ʊ] responses (crosses) for one subject as a function of stimulus number and vowel environment.

results in the #V#, [wVw], and [jVj] contexts for one subject (JH). To present the results in a fairly compact form, a quantification of the data has been attempted. On the assumption that the probability of making a given response increases (or decreases) in the region of a phoneme boundary, according to the function of an integrated normal distribution, ogives were fitted to the individual identification graphs as shown in Fig. 4. The mean value of the distribution corresponds to the boundary, or the 50% crossover point along the continuum. Its standard deviation is inversely related to the steepness of the cumulative curve. Mean values $\bar{x}$ and standard deviations σ for such distributions are presented in Table I for individual observers and for the three stimulus conditions #V#, [wVw], and [jVj]. The duration of the vowel segment was 200 msec in this material. To establish $\bar{x}$ and σ, the data were also plotted on normal-probability graph paper and a best-fitting straight-line approximation was determined in the region of the phoneme boundary by visual inspection.[12] This transformation of percentages into σ or z units has the effect of magnifying the importance of "noise" inherent in data points close to the asymptotes of the distribution, whereas errors of equal magnitude at more and more central values are given successively less weight. This fact was considered in the curve-fitting procedure by restricting it to the transitional segment between asymptotes.

Before considering the boundary locations, something should be said about the variability of the data. It

[12] A somewhat different procedure was used for the [w] data. It is described later in Sec. VI.

TABLE I. Fifty percent crossovers or boundary locations ($\bar{x}$) and standard deviations (σ) of identification functions fitted to the vowel responses obtained for steady-state condition ($\bar{x}_v$ and σ_v), the [w-w] context ($\bar{x}_w$ and σ_w) and the [j-j] context ($\bar{x}_j$ and σ_j). The numbers are in terms of stimulus numbers ($\bar{x}$) and continuum steps (σ). Duration of vowel stimuli: 200 msec.

Subject	$\bar{x}_v$	σ_v	$\bar{x}_w$	σ_w	$\bar{x}_j$	σ_j
CA	13.7	2.8	6.7	5.0	...	...
KE	12.7	2.3	9.0	3.5	14.5	3.7
FE	13.2	2.2	8.3	3.5	14.9	3.0
KR	12.9	2.4	9.6	4.7	14.4	2.4
SO	14.4	2.2	10.4	4.3	...	...
AD	13.0	2.6	12.3	4.0	14.3	4.0
JH	14.7	0.9	8.6	1.3	16.2	1.7
MS	14.8	0.8	13.2	1.7	11.8	1.2
DW	12.3	1.2	12.1	0.9	9.0	2.0
HW	14.0	1.3	17.1	4.2	11.0	3.3

TABLE II. Data from Table I. Displacement of boundaries for CVC stimuli in relation to #V# stimuli. Positive numbers indicate downward shifts, negative numbers upward shifts.

Subject	$\bar{x}_v - \bar{x}_w$	$\bar{x}_v - \bar{x}_j$
CA	7.0	...
KE	3.7	−1.8
FE	4.9	−1.7
KR	3.3	−1.5
SO	4.0	...
AD	0.7	−1.3
JH	6.1	−1.5
MS	1.6	3.0
DW	0.1	3.3
HW	−3.1	3.0
Median	3.5	−1.4

appears from Table I that the σ values are somewhat larger for the consonantal contexts than for the #V# situation. This result may be taken to indicate that subjects found it more difficult to make the judgments in the former cases. It is not unlikely that the consonantal stimuli struck some observers as more unnatural than the sustained vowels. For, though (by some stretch of the imagination) a [wVw] stimulus might roughly sound somewhat like *will* or *wool*, it is hard to find any English words to match the [jVj] stimuli. The difficulty of the consonant environments can also be inferred from the fact that three subjects failed altogether to show crossovers in the CVC situations, as did subjects CA and SO in the [j] environment. These data deviate from those of the others either in that the percentages of [ɪ] and [ʊ] answers fluctuate around 50% all along the stimulus continuum, or in that the subject favors a single response close to 100% of the time. Since cumulative ogives serve no descriptive purpose in such cases, these data are not tabulated. For the remainder of the material, clear crossovers were obtained. Yet some subjects showed uncertainty as to the identity of the vowel, even in clearly "asymptotic" regions of the stimulus continuum, and would, in such cases, produce asymptotes closer to 10% and 90% rather than to the expected 0% and 100%. It is, thus, clear that observers differed markedly in the case with which they made the identifications.

An examination of the boundary locations for the three experimental conditions indicates that for most of the subjects the [w] environment shifts the boundary towards low-stimulus numbers in relation to its position in the #V# situation. This general tendency is brought out most clearly in Table II where the distances between the boundaries for the consonant and the steady-state vowel contexts are presented for the observers individually and in terms of median values. The median boundary displacement for the [w] context is 3.5 continuum steps in the direction of low-stimulus numbers, and that for the [j] context is smaller (1.4 steps) but occurs in the opposite direction. These shifts, when expressed in terms of the frequency steps of the second formant, are approximately 185 cps ([w]) and 75 cps ([j]). Hence, there is a clear tendency for the boundary location to vary with context: it shifts in the direction of the consonantal loci. The [j] stimuli yield less consistent results in that the shifts occur towards both high- and low-stimulus numbers. The data of subject HW differ most markedly from the general trend of the material both in the direction and in the magnitude of the shifts.

In Fig. 4, the boundary shifts have the same directions as the median displacements discussed above. It is apparent that a larger displacement is obtained for the [w] context than for the [j] context. The [w] stimuli were those with convex frequency-time variation in the second and third formants, so that the initial transitional segment always contained positive or rising transitions. The midpoint vowel patterns at the termination of these segments invariably represented the frequency *maximum* of the formant in question (F_1 included). In terms of steps along the continuum, the boundary shift is about 320 cps in F_2. Describing this effect in terms of [ɪ] responses alone, it can be said that in 100% of the cases stimulus No. 11 in the [w] series appeared equivalent to stimulus No. 17 in the #V# series, as regard vowel identity. Thus, in the [w] context, the pattern of maximal extent of the formant transitions need only be about $F_1=350$, $F_2=1525$, and $F_3=2575$ for it to be consistently labeled [ɪ], whereas, in the #V# situation, a more "acute" pattern of $F_1=350$, $F_2=1845$, $F_3=2740$ is required to elicit equally consistent [ɪ] responses. The [j] stimuli had second and third formants that varied in a concave fashion as a function of time, so that the initial transitions were falling (except for the first formant transitions that were identical in the [w] and [j] series and exhibited a rise-fall or convex time variation). At the point of zero rate of change, formants (F_2 and F_3) reached their *minimum* frequency values. The boundary shift brought about by the [j] context is much smaller than for [w], only about 1.5 continuum steps as compared with 6 for [w]. On the whole, the identity changes

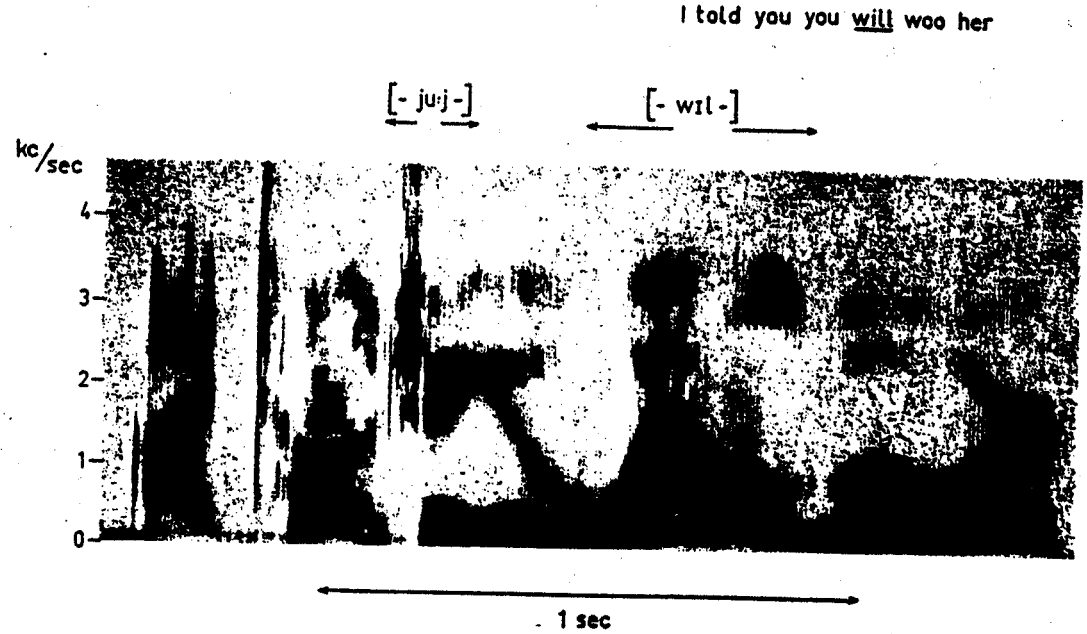

FIG. 5. Spectrogram of *I told you you will woo her* pronounced naturally with contrastive stress on *will* and reduced stress on the first *you*. Notice the small difference between the central F_2 values in [ju:j] and [wɪl]. Compare this utterance with that in Fig. 6.

are most extreme in the region of Vowel Patterns 11, 12, 13, and 14, where symbol assignment shifts radically from 100% [ɪ] responses in the [w] context to nearly 100% [ʊ] judgments in the [j] environment.

V. DISCUSSION

A. Complementarity of Vowel Production and Vowel Perception

The results of the identification experiment suggest that vowel stimuli that were identical with respect to midpoint formant frequencies, but differed with respect to formant transitions, were not perceptually equivalent. One purpose of the experiment was to investigate the relative rôles of formant transitions and such midpoint patterns in vowel recognition. The present results appear to indicate that the identity of a vowel stimulus is determined not only by the formant pattern at the point of closest approach to target, but also by the direction of the adjacent formant transitions. A second objective was to ascertain whether operations contributing to the reduction of ambiguity can be invoked in the recognition of monosyllabic nonsense speech. Do the particular effects observed in this experiment serve this purpose?

That, in fact, they do appears from the following considerations. As mentioned earlier, formant frequencies at the center of vowels may undershoot their target values and be displaced away from the values observed for steady-state conditions in the direction of adjacent consonantal loci. (Refs. 2, 3, 4, and 6.) In the production of syllables like [wɪw] and [wʊw], formants will, thus, be shifted towards their values for [w]; and for syllables like [jɪj] and [jʊj], towards their values for [j]. An illustration of such effects is given in Figs. 5 and 6, which show spectrograms of the utterances *I told you you will woo her* and *I told you you will woo her*, pronounced by a native male speaker of American English. The utterances contain the phonetic segments [−ju:j−] and [−wɪl−] under different conditions of stress that approximate some of the test syllables used. Table III shows differences between F_2 values for [ɪ] and [u:] in a "say hVd again" environment and the above segments. Each value is based on 10 measurements and has been sampled at points of closest approach to target. These data confirm the results of earlier investigations by demonstrating that formant undershoot occurs in the direction of the immediately adjacent loci and increases with shorter duration and decreasing prominence (Refs. 3 and 4). There is an asymmetry between the magnitudes of undershoot observed in the [j-j] and [w-l] environments, the displacement being greater in the [w-l] environment for both conditions of stress. This result may indicate a

TABLE III. Differences in cycles per second between the second formant values of [ɪ] and [u:] in "null" environments and context (F_2 undershoot).

	Stressed	Unstressed	Context
[u:]	0	−660	[j-j]
[ɪ]	+420	+970	[w-l]

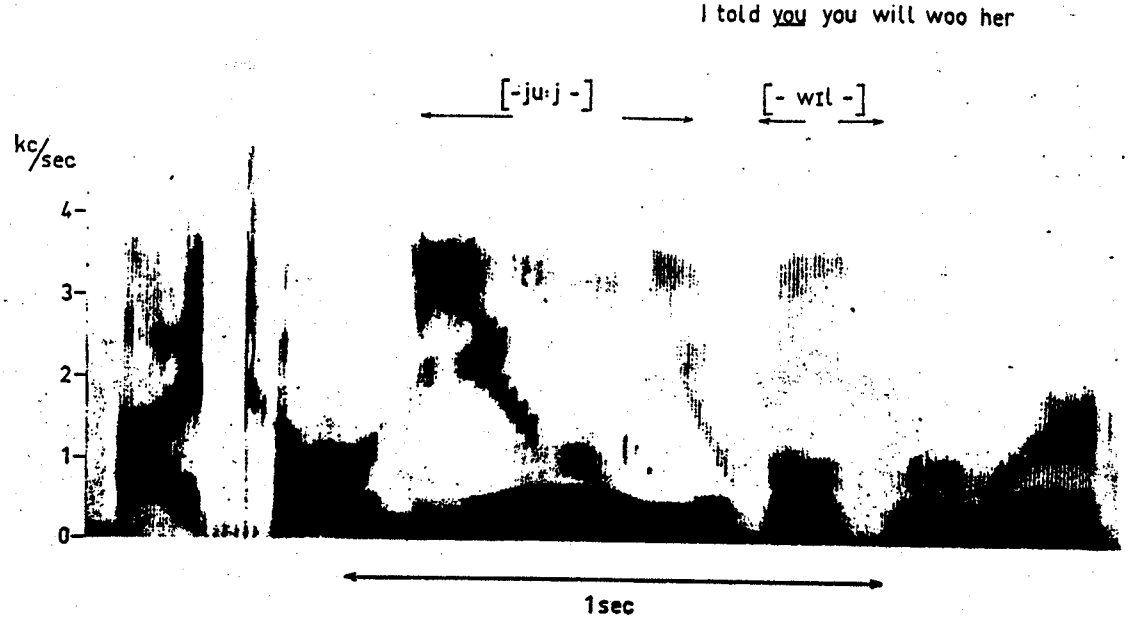

FIG. 6. Spectrogram of *I told you you will woo her* with contrastive stress on the first *you* and reduced stress on *will*. F_2 samples taken in the middle of the vowels of [ju:j] and [wɪl] are almost identical.

difference between the time constants of the movements involved, but may also be related to differences in locus-target distances, vowel duration, and stress.[12]

As noted earlier, the boundary shifts observed in Fig. 4 also occur in the direction of the loci. In terms of the stimulus continuum the [w] boundary is located six steps away from the position of the #V# boundary: It has been shifted toward low-stimulus numbers, that is, toward the [w] loci. Similarly, the [j] boundary has been slightly displaced toward the [j] loci. For this subject, it is clear that *categorization of the continuum is adjusted* in the different environments *so as to compensate for an undershoot effect* in the vowel stimuli. From the point of view of certain acoustic attributes of vowels, it would appear that the processes of vowel production and vowel recognition have complementary rôles: recognition compensates for production.

B. Active Recognition Model

The complementarity between production and recognition at the level of the nonsense syllable can be shown to be compatible with, for instance, an "active" perceptual analysis. By an active strategy, it is meant one that embodies an internal replication of the stimulus. The speech recognition model proposed by MacKay,[15,16] analysis by synthesis (Ref. 2),[17,18] and the motor theory[15,20] fall in this category. The steps by which the recognition of syllables like [jɪj], [wʊw], etc. might proceed according to such a scheme are as follows: The stimulus undergoes peripheral auditory analysis. A decision is made as to the contextual frame: Is it #V#, [wVw], or [jVj]? A computation is made of the auditory consequences of realizing an [ɪ] articulatorily and acoustically in the selected context. The result of this replication is compared with the input stimulus. The error is measured. The vowel [ʊ] is rep-

[12] An interesting question is to what extent the control of the articulatory goals underlying the production of American English vowels remains invariant under changes of stress. Current linguistic analysis of American English[14] postulates a vowel reduction rule that, in certain situations, would replace, for instance, an [ɪ] vowel by an [ɨ] provided that at least tertiary stress had been assigned to the [ɪ] previously in the stress cycle. Whether such a rule reflects a restructuring of the articulatory control of vowel pronunciation or merely describes an automatic physiological consequence of weak stress falls outside the scope of the present discussion.

[14] N. Chomsky and G. A. Miller, "Introduction to the Formal Analysis of Natural Languages," in *Handbook of Mathematical Psychology* (John Wiley & Sons Inc., New York, 1963), Vol. 2, pp. 269–321.

[15] D. M. MacKay, "Mindlike Behaviour in Artefacts," Brit. J. Sci. 2, 105–121 (1951).

[16] D. M. MacKay, "The 'Active/Passive' Controversy," paper presented at the Seminar on Speech Production and Perception, Leningrad 13–16 Aug. 1966 (to be published in Z. für Phonetik usw.).

[17] K. N. Stevens and M. Halle, "Remarks on Analysis by Synthesis and Distinctive Features," paper presented at the Symposium on Models for the Perception of Speech and Visual Form, Boston, Mass., 11–14 Nov. 1964 (to be published).

[18] K. N. Stevens and A. S. House, "Speech Perception," *Foundations of Modern Auditory Theory*, J. Tobias and E. Schubert, Eds. (to be published).

[19] A. M. Liberman, F. S. Cooper, K. S. Harris, and P. F. MacNeilage, "A Motor Theory of Speech Perception," in *Proceedings of the Speech Communication Seminar Stockholm 1962*, (Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, 1963), Vol. 2, paper D3.

[20] A. M. Liberman, F. S. Cooper, M. Studdert-Kennedy, K. S. Harris, and P. F. MacNeilage, "Some Observations on the Efficiency of Speech Sounds," paper presented at the Seminar on Speech Production and Perception, Leningrad, 13–16 Aug. 1966 (to be published in Z. für Phonetick usw.).

licated in a similar manner in the same context. Again the result is compared with the input. The vowel yielding the smaller error is the response. Accordingly, at the extreme low end of the vowel continuum, it is [ʊ] that gives the smaller error in the #V# context. As the stimulus number increases it is still [ʊ] that gives the better match until stimuli near the boundary are encountered. Here errors are about the same for both alternatives. It is seen from Fig. 4 that, in the #V# context, the boundary occurs, not at the midpoint of the continuum, but around Stimuli 14 and 15. This asymmetry might indicate that the listener in question assigns a more central continuum location to a typical [ʊ] sound than to an [ɪ] which appears to be located at the extreme high end. The same reasoning applies to the other contexts. However, if the process of internal replication is to simulate accurately the transformations from neural commands to sound, coarticulation will be one of the features replicated. Hence, it should predict different formant patterns for the two vowels in the various contexts. Not only will there be transitions in the case of the [w] and [j] environments and more or less steady-state formants for #V#, but the formant pattern at closest approach to target will also be different owing to the undershoot effect.

Consider a vowel formant pattern in the [wVw] context having an intermediate position on the stimulus continuum; Stimulus 11, for instance. The time variations of the second formant of this stimulus are shown to the left in Fig. 7. The ordinate has a mel scale. The internal regeneration mechanism supplies two alternative comparison patterns [wɪw] and [wʊw] (upper right of the figure). It is assumed that these patterns have been computed in accordance with the dynamic laws of speech production. Thus, the point of closest target approach falls short of the intended target frequencies for [ɪ] and [ʊ]. The assumptions underlying the construction of the replicated patterns are: (a) the

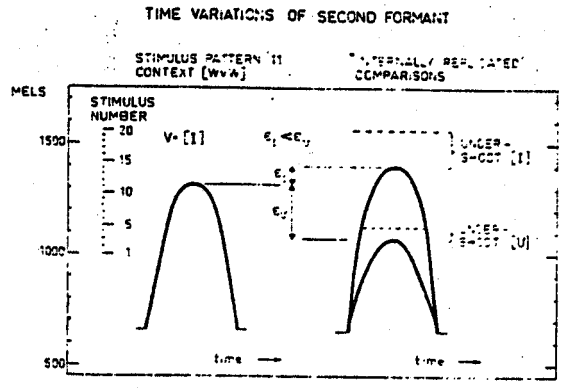

FIG. 7. "Active" analysis of Stimulus Pattern 11 in the [w-w] series (in terms of second formant only). Stimulus Pattern 11 is compared with two "internally computed" candidates for the best match: [wɪw] and [wʊw]. These hypothetical alternatives exhibit undershoot in relation to the targets (dashed). In this case, [wɪw] gives the smaller error ($\varepsilon_I < \varepsilon_U$), which makes [ɪ] the vowel response.

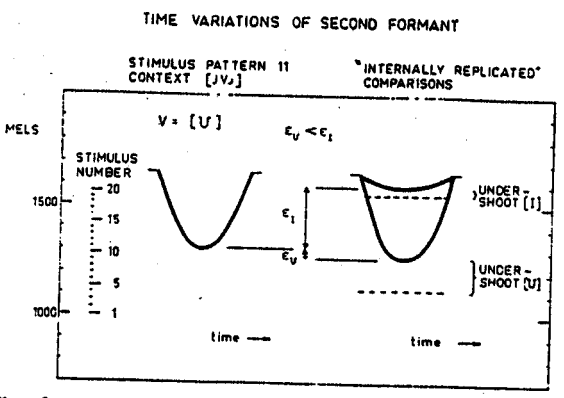

FIG. 8. Analogous to Fig. 7. Stimulus Pattern 11 occurs in the [j-j] context. Notice that it is [jʊj] that now gives the smaller error ($\varepsilon_U < \varepsilon_I$) so the response is [ʊ].

[ʊ] and [ɪ] target patterns correspond to Stimuli 4 and 20, respectively. ([ʊ] is given a slightly more central location than [ɪ]); (b) the undershoot in F_2 depends on the locus–target distance in question and is estimated at 25% of this distance for the present case. For the purposes of the present discussion, the deviation of the theoretically generated alternatives from the stimulus is taken to be defined simply as the difference in mels between these patterns at the instant of closest approach to target. If the mismatch for [ɪ], henceforth denoted ε_I, is larger than that for [ʊ], ε_U, the vowel recognized is [ʊ], and conversely; since in the present case of [wVw] $\varepsilon_I < \varepsilon_U$ the vowel response is [ɪ]. In Fig. 8 is shown the identical vowel formant configuration now in the [jVj] context. The internally generated candidates exhibit the same dynamic features as the previous patterns. It is seen that, although the formant configuration at zero rate of change is the same as before, a comparison of the errors in this context gives the opposite result. Since $\varepsilon_U < \varepsilon_I$, the vowel recognized is [ʊ]. As remarked above, such radical identity changes can be observed in Fig. 4. in the region of Vowel Patterns 11, 12, 13, and 14 for the two consonantal environments. For [wVw], ε_I is smaller than ε_U for stimulus numbers larger than 11, whereas the situation where $\varepsilon_I > \varepsilon_U$ obtains will be approached as lower stimulus numbers are encountered. Accordingly the boundary between [ɪ] and [ʊ] lies *below* Pattern 11 in the [w] environment. Similarly, in the case of [jVj], ε_U is smaller than ε_I for stimulus numbers below 11 and the [ɪ]-[ʊ] boundary, or the switchover from ($\varepsilon_U < \varepsilon_I$) to ($\varepsilon_U > \varepsilon_I$), occurs *above* Stimulus 11. As seen earlier in Fig. 4 and Table II, this is also the direction in which the experimentally observed boundary shifts occur. If the exact locations of the boundaries are taken to be found at a place where $\varepsilon_I = \varepsilon_U$ the expected boundaries occur between 8 and 9 (w), 12 and 13 (#V#), and 15 and 16 (j). The observed locations shown in Fig. 4 are about 8.5 (w), 14.5 (#V#), and 16 (j). The major discrepancy, two continuum steps, is obtained for

V#. The asymmetrical distribution of the estimated locations is in agreement with the data and is due to the definition of errors in terms of mels and the choice of target values for [ɪ] and [ʊ]. This asymmetry compares favorably with that observed in the formant-undershoot data from natural utterances similar to the present test syllables (Figs. 5 and 6).

VI. EXPERIMENTAL RESULTS II

A. Fast Stimuli (Vowel Duration: 100 msec)

The vowel recognition behavior under analysis is of great complexity. Any attempt to model it at present must be functional rather than structural. Accordingly, little can be said now about the actual mechanisms underlying the boundary shifts. On the other hand, the active model suggested above, although informal, provides a convenient starting-point for a further examination of the complementarity of production and perception at the nonsense syllable level. As used here, complementarity refers to the observation that, from the point of view of the phonetician, mechanisms of production introduce ambiguity at the acoustic level, whereas perceptual mechanisms resolve it. If the dependence of undershoot on the rate of talking (Refs. 3 and 4) were incorporated in the active recognition scheme, the internally predicted undershoot would depend on a measurement of the duration of the vowel. Since there is a general tendency in this material for the boundary shifts to occur in such a direction that they in fact compensate for undershoot in the vowels (whether subconsciously "assumed" or not by the listener), it is reasonable to ask whether a compression of the time scale of the same stimuli patterns would give rise to even larger boundary shifts. These considerations lead to a new experiment that might establish whether recognition at the nonsense-syllable level corrects for the variance and ambiguity in vowels to the same extent as they are introduced by production.

As mentioned earlier, in the Section on stimulus specification, two sets of stimuli were synthesized. One set contained vowels 200-msec long. The experimental results presented so far pertain to this set. During the listening sessions, subjects made responses also to a set of fast stimuli that differed from the slow ones only in that the time scale of the vowel segments had been compressed by a factor of 2 (Fig. 3). Vowel segments were consequently 100-msec long, and the over-all length of the CVC syllables was 140 msec. These sounds were presented under conditions described in the Section on experimental procedures. These conditions were identical for fast and slow stimuli, but the two types of stimuli occurred in separate groups on the test tapes. The data were plotted in the form of identification graphs, and ogives were fitted to the data points. The results of this experiment closely resemble those already discussed. Table IV shows crossover values ($\bar{x}$) and standard deviations (σ) for the three stimulus conditions. Boundary shifts occur as before (Table V): There are substantial downward shifts for [w] and upward shifts for [j]. Exceptions from this tendency are found among subjects who also deviated earlier. In similar agreement with the previous results, standard deviations are somewhat smaller in the # V# context, and Subjects CA and SO have trouble with the [j] stimuli, showing no crossovers.

TABLE IV. Fifty percent crossovers ($\bar{x}$) and standard deviations (σ) for data from a supplementary experiment in which vowel stimuli were 100-msec long but otherwise identical to those of Table I.

Subject	$\bar{x}_v$	σ_v	$\bar{x}_w$	σ_w	$\bar{x}_j$	σ_j
CA	12.6	2.5	6.0	3.1	...	...
KE	12.6	2.2	6.9	2.6	12.0	7.2
FE	12.1	2.2	5.5	3.4	16.2	2.6
KR	12.1	2.0	5.6	2.6	12.5	3.3
SO	13.4	1.8	9.0	1.2	...	...
AD	12.5	2.5	8.0	4.0	13.3	3.3
JH	14.1	1.0	8.5	1.5	16.9	1.4
MS	14.3	0.8	12.0	3.7	11.8	1.7
DW	11.3	1.2	11.5	1.2	8.8	2.3
HW	14.0	1.4	17.7	3.8	11.0	3.7

TABLE V. Data from Table IV. Displacement of boundaries for CVC stimuli in relation to # V# stimuli.

Subject	$\bar{x}_v - \bar{x}_w$	$\bar{x}_v - \bar{x}_j$
CA	6.6	...
KE	5.7	0.6
FE	6.6	−4.1
KR	6.5	−0.4
SO	4.4	...
AD	4.5	−0.8
JH	5.6	−2.8
MS	2.3	2.5
DW	−0.2	2.5
HW	−3.7	3.0
Median	5.0	0.1

Consider again the active model and the replicated patterns corresponding to [ɪ] in the [w] and [j] contexts. The duration of the vowels is now 100 msec. Let the internal replication mechanism make the same calculations as before, except that the undershoot effect is now assumed to be larger, say, 50% of the locus–target distance. As before, the [ɪ]–[ʊ] boundaries are located at a point on the continuum that yields equally large errors in terms of mels and the midpoint value of the second formant. These conditions now give boundary locations slightly above Stimulus No. 4 for [w], and close to 18 for [j]. The duration decrease of the stimuli should consequently cause a further shift of the boundaries by about four steps for [w] and by two steps for [j] from their previously estimated positions. Wherever a graphical curve-fitting procedure made a significance test of differences between the fast and slow data seem worthwhile, the 50% crossover

TABLE VI. Shifts of the [w-w] and [j-j] boundaries associated with a change of vowel duration in the stimuli. Positive numbers refer to upward shifts and conversely.

Subject	$\Delta\bar{x}_w$	$\Delta\bar{x}_j$
CA	−0.7	...
KE	−2.1***	−2.5
FE	−2.9***	+1.3
KR	−3.9***	−1.9
SO	−1.4*	...
AD	−4.3***	−1.0
JH	−0.1	+0.7
MS	−1.2***	0.0
DW	−0.6*	−0.2
HW	0.6	0.0

points, or means ($\bar{x}$), and the standard deviations (σ) were established by means of probit analysis.[21,22] According to this method, the best-fitting lines were determined after transforming response percentages to weighted values or probits, a procedure whereby points receive weights depending on both their inherent statistical reliability and their relative location in the distribution. Only the data lying between asymptotic regions were used. The method also provides an estimate of the variance of the crossover points, which makes it possible to use a t test to judge the significance of crossover differences. In this test, an infinite number of degrees of freedom was assumed.

Table VI shows the boundary displacement associated with a duration change in the stimuli. The values given refer to the differences between "fast" and "slow" boundary locations. A negative number thus indicates a downward shift, a positive number an upward shift. It is seen that for [w], shifts occur in the expected direction. Their magnitudes are, in general, smaller than estimated. On the other hand, the differences are probably not due to chance as indicated in the Table by the conventional asterisk notation. Only the [w] data seemed worth processing statistically. Thus, Tables I and IV show graphical estimates for #V# and [jVj], and computed estimates for [wVw]. In the case of [jVj], the results are less consistent with respect to the direction of the shifts that are small.

VII. PERIPHERAL REPRESENTATION OF VOWEL SOUNDS

The response of the human auditory system to a sound can be shown to depend appreciably on the acoustic context of this sound. Such dependence is exemplified for instance by adaptation and fatigue phenomena that manifest themselves as a time-varying short-term or more persistent reduction of auditory sensitivity.[23–25] Threshold shifts of this kind introduce a temporary decaying skewness in the frequency response of the ear and have for this reason been considered responsible for certain pitch-shift effects. It has been found, for instance, that the pitch attributed to pure tones tends to be displaced away from a frequency region whose threshold has been temporarily raised by preceding or simultaneous stimulation, towards a more sensitive region (Ref. 23).[26,27] Whereas there is a great deal of information in the literature on such contextual effects in the perception of pure tones, there appear to be few, if any, data on the peripheral representation of spectral shape in complex sounds like vowels and still fewer on the dependence of such representations on various environmental conditions. There is, however, a study by Brady *et al.*[28] on the perception of sounds with a rapidly varying formant frequency that may be of some relevance to the present discussion. In that investigation, listeners were asked to find a best match between a test sound and an adjustable comparison stimulus. Both stimuli were single-formant sounds of equal duration in the range of 20–50 msec and generated by identical periodic excitation. In the test stimulus, the resonant frequency varied linearly over a 500-cps range; the resonant frequency of the comparison stimulus was steady-state, but could be controlled by the observers. The results indicate that the preferred location of the formant frequency of the comparison stimulus was a value close to the terminal frequency of the varying formant. It is of interest to note that this tendency was stronger for a faster rate of frequency change and somewhat more pronounced for upward frequency ramps than for downward ones. In conclusion, the authors suggest that "there may be some overshoot or extrapolation in the processing of brief stimuli characterized by rapidly changing spectra." There are several possible ways in which these results could be interpreted. The foregoing discussion suggests one: The hypothetical overshoot mechanism (whose existence is supported by the tendency to place the comparison resonance still closer to the terminal frequency for a faster rate of frequency change) could tentatively be identified with adaptational processes.

[21] J. P. Guilford, *Psychometric Methods* (McGraw-Hill Book Co., New York, 1954).
[22] D. J. Finney, *Probit Analysis* (Cambridge University Press, Cambridge, England, 1947).
[23] G. von Békésy, *Experiments in Hearing* (McGraw-Hill Book Company, New York, 1960), pp. 366–368.
[24] R. Plomp, *Experiments on Tone Perception* (Institute for Perception RVO-TNO, Soesterberg, The Netherlands, 1966), pp. 20–22.
[25] E. Lüscher and J. Zwislocki, "The Decay of Sensation and the Remainder of Adaptation after Short Pure-Tone Impulses on the Ear," Acta Oto-Laryngol. 35, 428–445 (1947).
[26] J. P. Egan and D. R. Meyer, "Changes in Pitch of Tones of Low Frequency as a Function of the Pattern of Excitation Produced by a Band of Noise," J. Acoust. Soc. Am. 22, 827–833 (1950).
[27] J. C. Webster and E. D. Schubert, "Pitch Shifts Accompanying Certain Auditory Threshold Shifts," J. Acoust. Soc. Am. 26, 754–758 (1954).
[28] P. T. Brady, A. S. House, and K. N. Stevens, "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency," J. Acoust. Soc. Am. 33, 1357–1362 (1961).

During the test stimulus, auditory events can be grossly pictured as follows: There occurs a continuous change in the sensitivity of the ear owing to adaptation. At any given moment the "tracking" of the formant peak may be shifted towards a region of lower threshold, that is, upwards for an ascending ramp and downwards for descending ramps, the amount of displacement depending on how far adaptation has progressed at that moment. This effect would accordingly be one of the factors contributing to shifting matches closer to the terminal frequencies of the ramps. If verified by future experimentation[29] these speculations invite the conclusion that the peripheral auditory representation of spectral formant peaks depends on the rate of formant frequency change and offer a possible explanation of some of the present rate- and context-dependent boundary shifts. According to this view, vowel production and vowel perception could be said to be complementary in the sense that articulatory activity is characterized by undershoot and perception by overshoot.

VIII. DISTINCTIVE FEATURES

There are various other ways of describing the results obtained, for instance, in terms of distincitive features. If, in a sequence of feature complexes, one of the features takes the values of +−+ in one case and −−− in another, the phonetic correlates of the middle segment need exhibit less pronounced "negativity" in absolute terms in the positive context than in the negative one. In Fig. 4, this is illustrated by the fact that the observer permits [ʊ] vowels to possess less marked cues of flatness (labialization) and graveness (velarization) in the plain (unrounded) and acute (palatalized) context of [j], and conversely [ɪ] vowels to present less trace of plain and acute attributes in the flat and grave environment of [w]. This interpretation exemplifies the statement that phonetic features are relational. Their values are specified along scales in relative terms.[30,31] The data may also be examined with respect to their implications about a talker's encoding strategy. In order to produce [ɪ] and [ʊ] sounds that remain distinctive, what is minimally required of him? Suppose that he produces [ɪ] and [ʊ] vowels that are most intelligible when their formant patterns approach the upper and lower ends, respectively, of the vowel continuum. In the [wVw] context, it is not necessary for an [ɪ] vowel to have a formant pattern that lies close to the upper end of the continuum. For example, F_2 and F_3 at 1525 and 2575 cps will be sufficient to elicit the correct response from the listener of Fig. 4. The situation is reversed, but similar, for the [jVj] environment and an [ʊ] vowel. In terms of the extent of formant movements in syllables like [wɪw] and [jʊj], this, consequently, means that a complete transition from loci to vowel target is not required. The transitions could undershoot their target frequencies for the vowel to a certain extent without distinctiveness being lost. Under such circumstances, it is not necessary for a talker to compensate for the sluggish dynamics of his articulatory mechanism by reorganizing his control of it in such a way that undershoot is avoided. The present findings suggest that there may be recognition processes that compensate for this sluggishness and absence of reorganization. Not only do we "speak to be heard in order to be understood," but we obviously also listen to hear in order to understand.

IX. RÔLE OF EXPECTANCY

It might be argued that the boundary shifts have to do with differences in the predictability of the vowels in the different contexts. According to this view the more probable vowel could be given as a response to stimuli located in an intermediate uncertainty region between clear [ɪ] and [ʊ] alternatives. This bias should have the effect of moving the boundary towards the less probable vowel. It can be objected that such a mechanism would remain insensitive to a tempo change. Since duration-dependent shifts were observed, this view is difficult to defend. Moreover, it is not clear what should be meant by the notion of "more probable vowel" in this connection.

X. PRODUCTION AND PERCEPTION OF DIPHTHONGS

There is an old observation on the perception of diphthongs that appears pertinent in connection with the present results. Jespersen[32] writes: "Fallende Diphthonge. Hier entscheidet oft bloss die Richtung der Bewegung den resultierenden Laut. Statt dass man z.B. in beabsichtigtem [ɑi] den ganzen Weg von [ɑ] bis [i] geht, begnügt man sich damit, nur ein Stück zu gehen, indem das Ohr leicht *getäuscht wird* (italicized by us, BL/MSK) und die Phantasie leicht das Fehlende ergänzt." Similar remarks can be found in the works of other authors.[33] Thus, Jones states that "the English diphthong *ai* is one which begins at *a* and moves in the direction of *i*. *To give the right effect* (italics ours, BL/MSK) it is not necessary that *i* should be quite reached; the diphthong may, and generally does, end at an opener vowel than this, such as a fairly open variety of *e*." The italicized passages clearly imply that

[29] It is believed that information relevant to these questions can be obtained by means of standard psychoacoustical experimental techniques. Several projects designed to elucidate these problems are at present under way.

[30] F. de Saussure, *Cours de linguistique générale* (C. Bally & Sechehaye, Paris, 1916), pp. 1–337.

[31] R. Jakobson and M. Halle, *Fundamentals of Language* (Mouton and Company, 's-Gravenhage, The Netherlands, 1956), pp. 1–87.

[32] O. Jespersen, *Lehrbuch der Phonetik* (Verlag von B. G. Teubner, Leipzig, Germany, 1926), p. 208.

[33] D. Jones, *An Outline of English Phonetics* (W. Heffer & Sons Ltd., Cambridge, England, 1956), pp. 58–59 Sec. 224.

the auditory value of the second element is [i], in the opinion of these phoneticians. Thus, an articulatory movement [æe] or [æɛ] is heard as [ai] by the naïve listener. In terms of the second formant of such sounds, there would be a transition that would have a positive slope and would fall short of an [i] value, thus, terminating at a lower frequency. In spite of this reduced extent, an [i] element is claimed to be heard that is analogous to what was found for [wɪw] in the present material.

XI. SOME PARALLEL OBSERVATIONS

The results of Fujimura and Ochiai (Ref. 9) on vowel identification also bear on our own findings.[34] These investigators gated out 50-msec portions of vowels from Japanese words and made an analysis of listener responses to these segments presented in isolation. In general, the confusions could be explained in terms of coarticulation. Thus, it was found that an [u] from /yuyusii/ was recognized as /i/. If more of the context or the entire word had been presented, confusions would not have occurred to the same extent. The implication of this work (and of our own) is that the assignment of symbols to vowels normally involves some sort of context-sensitive routine. Whether these operations are of short-term (syllabic) and/or long-term (word) nature cannot be inferred from this Japanese investigation. The present findings on the categorization of the [ɪ]–[ʊ] continuum, on the other hand, do not exclude the interpretation that, among other processes, a short-term mechanism might be involved.

In agreement with the present results, Stevens[35] found that listeners tend to categorize a given vowel continuum differently depending on whether the vowels are isolated and steady-state, or embedded in a CVC frame. It is of interest to note that, the boundary shift observed also occurred in a direction so as to compensate for potential undershoot effects.

Fry *et al.*[36] observed a dependence of vowel labeling on context. These investigators asked listeners to identify synthetic steady-state vowels presented in ABX groupings. They conclude that the effect of sequence on vowel identification is considerable. There was a tendency for a given vowel stimulus along an [ɪ]–[æe] continuum to be judged as closer (more [ɪ] like) when paired with a more open sounding ([æe] like) vowel, and conversely, as more [æe] like in the neighborhood of a closer sounding stimulus. Similar context dependence in vowel perception was reported also by Ladefoged and Broadbent.[37] This study demonstrated that the vowel in a monosyllable could be influenced by the formant patterns used in a preceding carrier sentence. The results of Fry *et al.* appear to indicate that a shift of reference frame in vowel perception may occur although the stimuli do not contain formant transitions and are separated in time by as much as one second (Ref. 36). At present, it cannot be determined whether these contrast effects and the present analogous [ɪ]–[ʊ] boundary shifts are attributable to the same underlying mechanism. It is worth reiterating, however, that mechanisms of perceptual analysis whose operations contribute to enhancing contrast in the above-mentioned sense are precisely the type of mechanisms that seem well suited to their purpose given the fact that the slurred and sluggish manner in which human speech sound stimuli are often generated tends to reduce rather than sharpen such contrast.

XII. SUMMARY AND CONCLUSIONS

1. Listeners were found to categorize a vowel continuum ranging from [ɪ] to [ʊ] differently depending upon whether the context of the vowels was #V#, [w-w], or [j-j].

2. The locations of the boundary between [ɪ] and [ʊ] observed for [w-w] and [j-j] tended to be displaced towards the consonant loci in relation to its position for #V#.

3. The boundary shift was most marked for [w-w].

4. A decrease of the duration of the vowel stimuli gave results similar to those mentioned above but was associated with a slightly larger boundary displacement for [w-w].

There emerges from these two experiments the tentative conclusion that, in the recognition of monosyllabic nonsense speech, the identity of a vowel sound is determined not only by the formant pattern at the point of closest approach to target but also by the *direction* and *rate* of adjacent formant transitions. Boundary shifts in the [w] context occurred in such a direction as to compensate for formant-frequency undershoot in the vowels. Vowel recognition thus compensated for vowel production. In this sense, these processes were found to exhibit complementarity.

ACKNOWLEDGMENTS

We would like to express our deep gratitude to Dr. F. S. Cooper, Haskins Laboratories, New York, to Professor G. Fant, Department of Speech Communication, KTH, Stockholm, and to Professor K. N. Stevens, Massachusetts Institute of Technology, Cambridge,

[34] We wish to express our gratitude to Professors Fujimura and Ochiai for putting a prepublication draft of this paper at our disposal.

[35] K. N. Stevens, "On the Relations between Speech Movements and Speech Perception," paper presented at the Seminar on Speech Production and Perception, Leningrad, 13–16 Aug. 1966 (to be published in Z. für Phonetik usw).

[36] D. B. Fry, A. S. Abramson, P. D. Eimas, and A. M. Liberman, "The Identification and Discrimination of Synthetic Vowels," Language and Speech 5, 171–189 (1962).

[37] P. Ladefoged and D. E. Broadbent, "Information Conveyed by Vowels," J. Acoust. Soc. Am. 39, 98–104 (1957).

Massachusetts, for making it possible for us to do the present research and for their generous support and stimulating interest during all phases of it. We are also greatly indebted to Mrs. S. Felicetti for preparing the manuscript and for editorial aid, to Mr. Å. Florén for technical assistance, and to all other colleagues at the Department of Speech Communication, KTH, and at Haskins Laboratories who helped us in various ways. The cooperation of Professor W. L. Henke and Professor A. S. House during the exploratory stages of this work is gratefully acknowledged. To a considerable extent, the present investigation owes it existence to Professor A. M. Liberman who encouraged us all along and with whom we have had frequent and fruitful discussions. Thanks are due also to Dr. O. Franzén, Professor J. M. Heinz, Dr. A. W. Slawson, and Dr. S. E. G. Öhman for responding with many valuable thoughts and suggestions.

This research was supported by funds for speech research at Haskins Laboratories, at Research Laboratory of Electronics, MIT, and at the Department of Speech Communication, KTH.[38]

[38] The Joint Services Electronics Program; The National Science Foundation; the National Aeronautics and Space Administration; the U. S. Air Force Cambridge Research Laboratories; the National Institutes of Child Health and Human Development, the National Institutes of Health, U. S. Department of Health, Education, and Welfare; and the Swedish Technical Research Council (Statens Tekniska Forskingsråd).